Algoritma *Naive Bayes Classifier* Dan *Chi Squared Statistic* Untuk Analisis Sentiment *Tweets* Bahasa Indonesia dan Sunda

Yono Cahyono[1], Saprudin[2]
Teknik Informatika Universitas Pamulang
e-mail : [1]dosen00843@unpam.com, [2]dosen00845@unpam.ac.id

ABSTRAK

Perkembangan dalam penggunaan media sosial saat ini terutama di Indonesia sangatlah pesat. Negara Indonesia merupakan negara dengan berbagai aneka ragam salah satunya bahasa daerah. Masyarakat Indonesia dalam komunikasi sehari-harinya selain menggunakan bahasa Indonesia, ada sebagian masyarakat terutama yang tinggal di daerah Jawa Barat masih menggunakan bahasa daerah yaitu bahasa Sunda untuk menyampaikan pendapat, komentar, saran maupun kritik dan lain-lain di media sosial. Data dari media sosial ini dapat digunakan untuk menggali informasi sehingga dapat digunakan untuk pengambilan keputusan baik bagi individu maupun organisasi. Jumlah data dari media sosial yang sangatlah besar membuat manusia tidak dapat menganalisisnya secara manual. Analisis sentimen ini merupakan suatu proses mengklasifikasikan, menganalisis, mengevaluasi, baik pendapat, komentar, saran maupun kritik dan lain-lain, terhadap objek tertentu seperti individu, organisasi, peristiwa, produk atau layanan, untuk mendapatkan informasi. Algoritma klasifikasi *Naïve Bayes Classifier* (NBC) dan metode pemilihan fitur *Chi Squared Statistics* digunakan dalam proses analisis sentiment pada *tweets* berbahasa Indonesia dan berbahasa Sunda di media sosial *Twitter*, dengan dikelompokan ke dalam kategori positif, negatif dan netral. Proses klasifikasi *Naïve Bayes Classifier* dan metode pemilihan fitur *Chi Square Statistic* dapat mengurangi fitur yang tidak relevan dalam proses klasifikasi pada *tweets* berbahasa Indonesia dan Sunda dengan akurasi sebesar 81,25%.

Kata kunci: *Twitter*, Analisis Sentiment, Berbahasa Indonesia dan Sunda, *Naïve Bayes Classifier* (NBC), *Chi Squared Statistic*

ABSTRACT

The development in the use of social media at this time, especially in Indonesia is very rapid. The country of Indonesia is a country with a variety of diverse one of them regional languages. Indonesian people in their daily communication besides using Indonesian, there are some people, especially those who live in West Java, still use local languages, namely Sundanese to express opinions, comments, suggestions and criticisms and others on social media. Data from social media can be used to gather information so that it can be used for decision making both for individuals and organizations. The huge amount of data from social media makes it impossible for humans to analyze it manually. Sentiment analysis is a process of classifying, analyzing, evaluating, both opinions, comments, suggestions and criticisms, etc., on certain objects such as individuals, organizations, events, products or services, to obtain information. The Naïve Bayes Classifier (NBC) classification algorithm and the Chi Squared Statistics feature selection method are used in the sentiment analysis process in Indonesian and Sundanese language tweets on Twitter social media, grouped into positive, negative and neutral categories. The Naïve Bayes Classifier classification process and the Chi Square Statistics feature selection method can reduce irrelevant features in the classification process in Indonesian and Sundanese tweets with an accuracy of 81.25%.

Keywords **:** *Sentiment Analysis, Indonesian and Sundanese Language, Naïve Bayes Classifier (NBC), Chi Squared Statistic*

PENDAHULUAN

Pengguna media sosial saat ini sangatlah pesat khususnya media sosial *Twitter*. Negara Indonesia adalah negara dengan berbagai aneka ragam salah satunya bahasa daerah. Masyarakat Indonesia dalam berkomunikasi selain menggunakan bahasa Indonesia, ada juga sebagian masyarakat terutama yang tinggal di daerah Jawa Barat mereka menggunakan bahasa daerah yaitu bahasa Sunda untuk menyampaikan pendapat, komentar, saran maupun kritik dan lain-lain di media sosial. Perkembangan dalam penggunaan media sosial *online* contohnya seperti media sosial *Twitter*, menyebabkan keinginan untuk menggali informasi yang ada didalamnya. Pada media sosial *Twitter* terdapat istilah *tweet* yang merupakan suatu pesan atau status yang dibuat oleh penggunanya. Suatu *tweet* dapat mengekspresikan suatu perasaan atau keadaan dari pengguna *Twiter*. *Tweet* bisa mengandung sebuah opini atau pendapat dari penggunanya mengenai kejadian yang dialaminya. Opini atau pendapat tersebut dapat dimanfaatkan untuk penilaian baik bagi perorangan ataupun bagi perusahaan.

Twitter telah banyak digunakan oleh instansi pemerintahan, perusahaan dan perorangan sebagai media komunikasi. Tapi untuk menentukan dan memilah apakah suatu *tweet* tersebut dapat mengandung sebuah opini positif, negatif atau netral tentu bukan hal yang mudah jika *tweet* yang diteliti jumlahnya sangat banyak. Maka dengan permasalahan tersebut dapat di bangun sebuah sistem yang dapat melakukan analisis sentimen.

Beberapa teknik klasifikasi yang sudah banyak digunakan dalam proses klasifikasi data salah satunya adalah metode *Naïve Bayes* atau sering disebut dengan *Naïve Bayes Classifier* (NBC). Algoritma *Naïve Bayes* dipilih disebabkan karena algoritma ini sangat baik atau cocok untuk *short* data *text*. Kelebihan lain dari *Naïve Bayes Classifier* adalah metode ini sederhana tetapi memiliki akurasi dan performansi yang tinggi dalam proses klasifikasi teks [1] (Routray, 2013).

Sedangkan untuk seleksi fitur merupakan proses optimasi untuk mengurangi suatu set besar fitur dari sumber aslinya, untuk memperoleh sejumlah subset fitur yang relatif kecil dan signifikan untuk meningkatkan akurasi dalam proses klasifikasi.

Dalam penelitian ini menggunakan penggabungan metode *Naïve Bayes Classifier* (NBC) dan pemilihan fitur *Chi Squared Statistic* untuk analisis sentiment *tweets* bahasa Indonesia dan bahasa sunda.

BAHAN DAN METODE

Pengumpulan Data *(Dataset)*

Data yang dipakai dapam proses s*entiment analysis* ini diperoleh dengan cara *crawl* (mengumpulkan) data dari media sosial *Twitter***.** Media sosial *Twitter* dipilih mengingat pengguna media sosial *Twitter* saat ini merupakan salah satu media sosial yang banyak digunakan dan populer untuk digunakan dalam mengungkapkan opini atau pendapat mengenai sesuatu hal.

Tahapan berikutnya adalah melakukan proses awal terhadap dokumen atau *Pre-processing.*

Pre-Processing

Pre-processing (pemrosesan awal dokumen) merupakan tahapan awal digunakan untuk mentransformasikan dokumen ke dalam bentuk representasi yang lain. Tujuan dari tahapan ini adalah untuk mempermudah dalam proses pencarian *query* ke dalam dokumen, mempercepat dalam pemrosesan terhadap dokumen, dan mempermudah dalam proses mengurutkan dokumen-dokumen yang diambil (*retrieved*). Tahapan proses yang dilakukan dalam *pre-processing* adalah *casefolding, tokenize* dan *stopword removal* [2] (Berry & Kogan, 2010).

a. Case folding

Pada tahap ini dilakukan untuk pengubah huruf besar dalam dokumen menjadi huruf kecil semua. Yang diterima pada proses ini hanyalah huruf „a" sampai

dengan „z". Sedangkan karakter lain selain huruf dianggap sebagai *delimiter* dan dihilangkan.

b. Tokenizing

Pada tahap ini dilakukan setelah data uji melewati tahapan *Case Folding*. Dimana *Tokenizing* ini merupakan proses pemotongan *string input* berdasarkan tiap kata yang menyusunya, serta membedakan karakter-karakter tertentu yang dapat diperlakukan sebagai pemisah kata atau bukan.

c. Stopwords removal

Selanjutnya untuk tahapan *filter stopwords (dictionary)* adalah akan menghilangkan kata-kata pada daftar *stopwords (dictionary)* yang tidak memiliki arti.

Seleksi Fitur *Chi Square Statistic*

Pada penelitian ini menggunakan seleksi fitur *Chi Square Statistic*. Dengan cara menghitung nilai seleksi fitur *Chi Square Statistic* dengan persamaan sebagai berikut [3] (Yang & Pedersen, 1997) :

$$X^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (1)$$

Keterangan:

A : Banyaknya dokumen dalam kategori *c* yang mengandung *term t*

B : Banyaknya dokumen yang bukan kategori *c* tetapi mengandung *term t*

C : Banyaknya dokumen dalam katgori *c* tetapi tidak mengandung *term t*

D : Banyaknya dokumen yang bukan kategori *c* dan tidak mengandung *term t*

N : Total keseluruhan dokumen

Seleksi fitur *Chi Square Statistic* digunakan untuk melakukan kesesuaian pegamatan (*goodness of fit*) dari kategori dengan *terms*. Uji *Chi Square Statistic* dalam statistika diterapkan untuk menguji independensi dari dua peristiwa. Sedangkan dalam seleksi fitur berdasarkan teori statistika, dua peristiwa tersebut di antaranya adalah kemunculan fitur dan kemunculan kategori.

Tahap *Cross Validation*

Dalam tahap *cross-validation*, setiap *record* akan digunakan beberapa kali yaitu digunakan untuk data *training* dan untuk data *testing*. Untuk dapat mengilustrasikan pada metode ini, anggaplah data akan dipartisi ke dalam dua buah *subset*. Yang pertama dipilih satu *subset* tersebut untuk data *training* dan satu lagi untuk data *testing*. Kemudian akan dilakukan pertukaran fungsi dari *subset* sehingga *subset* yang sebelumnya sebagai *training set* akan menjadi *testing set* demikian juga sebaliknya.

10-fold cross-validation akan melakukan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian. Metode ini merupakan evaluasi standar yaitu *stratified 10-fold cross-validation* karena menunjukkan bahwa *10-fold cross-validation* adalah merupakan pilihan terbaik untuk mendapatkan hasil validasi yang lebih akurat. Keuntungan dari metode ini adalah menghindari *overlapping* pada data *testing*. *Test set* bersifat *mutually exclusive* dan secara efektif mencakup keseluruhan *data set*. Namun kekurangan dari pendekatan ini adalah banyaknya proses komputasi untuk melakuan pengulangan prosedur sebanyak *N* kali [4] (Gorunescu, 2011).

Tahap Klasifikasi

Pada tahap klasifikasi menggunakan algoritma *Naive Bayes Classifier* yang merupakan proses klasifikasi dengan metode probabilitas, yaitu memprediksi peluang pada masa yang akan datang, berdasarkan pengalaman pada masa lalu sehingga dikenal sebagai *teorema Bayes*. Algoritma *naive bayes classifier* merupakan algoritma yang digunakan untuk mencari nilai probabiitas tertinggi untuk mengklasifikasikan data uji pada kategori yang paling tepat [5] (Feldman & Sanger, 2007).

Evaluasi

Evaluasi performansi dilakukan untuk menguji hasil dari proses klasifikasi dengan cara mengukur nilai performansi dari sistem yang telah dibuat. Parameter pengujian yang digunakan untuk evaluasi adalah diperoleh dari tabel *Confussion Matrix* untuk perhitungan tingkat akurasinya,

Tabel 1 *Confussion Matrix*

	***true* Negatif**	***true* Positif**	***true* Netral**
***pred.* Negatif**	TN	FP	FNet
***pred.* Positif**	FN	TP	FNet
***pred.* Netral**	FN	FP	TNet

Keterangan pada Tabel 1, *True* Positif (TP) merupakan tupel positif di *dataset* yang diklasifikasikan positif, sedangkan *False* Positif (FP) adalah tupel positif di *dataset* yang diklasifikasikan negatif atau netral. *True* Negatif (TN) merupakan tupel negatif di *dataset* yang diklasifikasikan negatif, sedangkan *False* Negatif (FN) merupakan jumlah tupel negatif di *dataset* yang diklasfikasikan positif atau netral. *True* Netral (TNet) merupakan tupel netral di *dataset* yang diklasifikasikan netral, sedangkan *False* Netral (FNet) adalah tupel netral di *dataset* yang diklasifikasikan positif atau negatif.

Dari tabel *Confussion Matrix* selanjutnya dapat menghitung *accuracy, Recall Positives, Recall Negatives, Recall Neutral, Positives Predicted Value (PPV), Negatives Predicted Value (NPV), Neutral Predicted Value (NetPV).*

$$Accuracy = \frac{TP + TN + TNet}{TP + TN + FP + FN + FNet + TNet} \quad (2)$$

$$Recall\ Positives = \frac{TP}{TP + FN + FNet} \quad (3)$$

$$Recall\ Negatives = \frac{TN}{TN + FP + FNet} \quad (4)$$

$$Recall\ Neutral = \frac{TNet}{TNet + FP + FN} \quad (5)$$

$$PPV = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ \ Positives} \quad (6)$$

$$NPV = \frac{Number\ of\ True\ Negatives}{Number\ of\ True\ Negatives\ +\ Number\ of\ False\ \ Negatives} \quad (7)$$

$$NetPV = \frac{Number\ of\ True\ Neutral}{Number\ of\ True\ Neutral\ +\ Number\ of\ False\ \ Neutral} \quad (8)$$

Recall Positives (perolehan positif) adalah jumlah kasus dengan perolehan positif, *Recall Negatives* (perolehan negatif) adalah jumlah kasus dengan perolehan negaitif, dan *Recall Neutral* (perolehan netral) adalah jumlah kasus dengan perolehan netral. , *NPV* (nilai prediktif negatif) adalah jumlah kasus dengan hasil diagnosa negatif, *PPV* (nilai prediktif positif) adalah jumlah kasus dengan hasil diagnosa positif dan *NetPV* (nilai prediktif netral) adalah jumlah kasus dengan hasil diagnosa netral.

HASIL

Pengumpulan Data

Data *tweets* ini diperoleh dengan membuat program *crawling* dengan menggunakan *tools Anaconda.* Proses *crawling* secara otomatis akan mengambil data *tweets* yang mengandung kata "PERSIB".

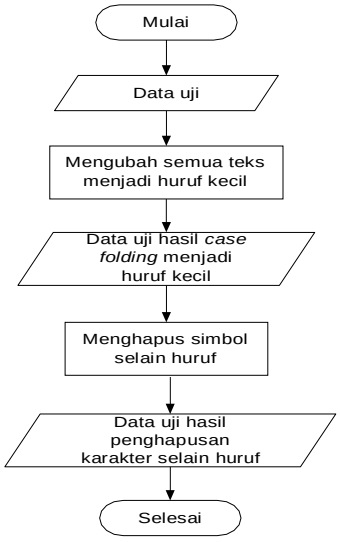
Sebanyak 640 data *tweets* yang akan diklasifikasikan ke dalam tiga kategori, kategori sentimen positif, kategori sentimen negatif dan kategori sentimen netral, yaitu dengan cara melakukan pelabelan secara manual terhadap data *tweets.*

Pre-*Processing*

Proses-proses yang dilakukan dalam pre-*processing* yaitu *casefolding*, *tokenize* dan *stopword removal.*

a. Case folding

Tahapan *case folding* mempunyai alur yang digambarkan pada gambar 1 sebagai berikut.



Gambar 1 *Flowchart* Tahapan *Case Folding*

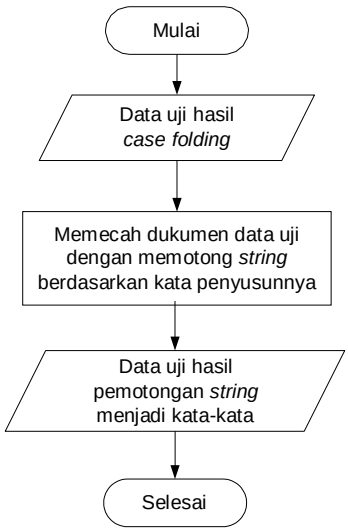
Contoh *case folding* data uji pada tabel 2 sebagai berikut:

Tabel 2 Contoh *Case Folding*

Data uji	**Hasil *case folding***
Tetep masih dukung persib, Erek susah erek seneng tetep persib	tetep masih dukung persib, erek susah erek seneng tetep persib

b. Tokenizing

Proses *tokenizing* ini mempunyai alur yang digambarkan pada gambar 2 sebagai berikut.



Gambar 2 *Flowchart* Tahapan *Tokenizing*

Contoh *Tokenizing* data uji pada tabel 3 sebagai berikut :

Tabel 3 Contoh *Tokenizing*

Data latih hasil *case folding*	**Hasil *tokenizing***
tetep masih dukung persib, erek susah erek seneng tetep persib	tetep masih dukung persib erek susah erek seneng tetep persib

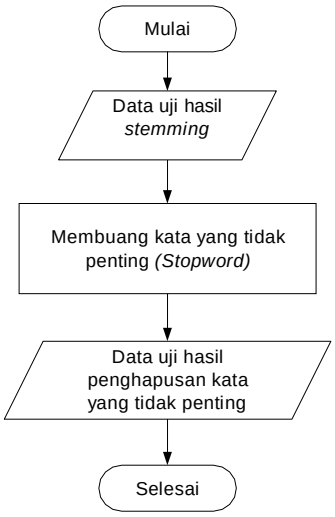
c. Stopwords removal

Selanjutnya untuk tahap *filter stopwords (dictionary)* adalah akan menghilangkan atau menghapus kata-kata yang tidak memiliki arti, dimana kata-kata tersebut terdapat pada daftar *stopwords* bahasa Indonesia dan bahasa sunda terdiri dari 999 kata-kata. Contoh kata-kata *stopwords* bahasa Indonesia dan bahasa sunda ditunjukan pada Tabel 4.

Tabel 4 Kata-kata S*topwords* Bahasa Indonesia dan Bahasa Sunda

ada	berapa	anu	ayeuna
adalah	berapakah	anyar	badag
adanya	berapalah	apa	bade
adapun	berapapun	aranjeun	bagean
agak	berarti	arek	baheula
agaknya	berawal	asa	bakal
agar	berbagai	atanapi	baruk

Proses *stopword (dictionary)* ini mempunyai alur yang digambarkan pada gambar 3 sebagai berikut.



Gambar 3 *Flowchart Stopword*

Contoh *stopword (dictionary)* data uji pada tabel 4 sebagai berikut.

Tabel 4 Contoh *Stopword (dictionary)*

Data latih hasil *tokenizing*	**Tahapan *stopword***
tetep	tetep
masih	dukung
dukung	persib
persib	erek
erek	susah
susah	erek
erek	seneng
seneng	tetep
tetep	persib
persib	

Seleksi Fitur *Chi Square Statistic*

Seleksi fitur *Chi Square Statistic* digunakan untuk pegamatan kesesuaian (*goodness of fit*) dari kategori dengan *terms*. Uji *Chi Square Statistic* dalam statistika diterapkan untuk menguji independensi dari dua peristiwa. Sedangkan dalam seleksi fitur berdasarkan teori statistika, dua peristiwa tersebut di antaranya adalah, kemunculan dari fitur dan kemunculan dari kategori.

Misalkan untuk menghitung nilai seleksi fitur *Chi Square Statistic* dari kategori *(c)* "Positif" dengan *term (t)* "dukung".

Tabel 5 Perhitungan Seleksi Fitur *Chi Square Statistic*

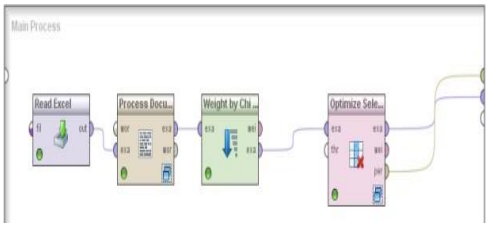
Kategori	**Jumlah dokumen**	***Term (t)* "dukung"**
Positif	91	3
Negatif	135	1
Netral	414	0
Total	**640**	**4**

Berdasarkan tabel diatas maka diperoleh A = 3, B = 1, C = 88, D = 548 dan N = 640, lakukan perhitungan nilai seleksi fitur *Chi Square Statistic* dengan persamaan (1) sebagai berikut :

$$X^2(t,c) = \frac{640 \times ((3 \times 548) - (88 \times 1))^2}{(3 + 88) \times (1 + 548) \times (3 + 1) \times (88 + 548)} = 16{,}29 \quad (9)$$

diperoleh bobot nilai 16,29 untuk *term (t)* "dukung" dengan kategori (c) "Positif". Agar nilainya dapat dibandingkan secara langsung antara *term* dengan kategori yang sama maka dilakukan normalisasi.

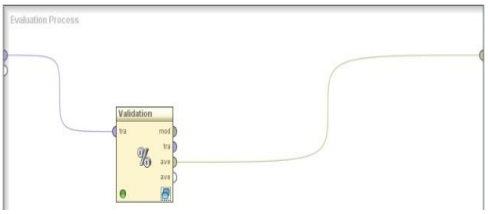
Proses klasifikasi dengan seleksi fitur *Chi Square Statistic* ditunjukan pada Gambar 4.



Gambar 4 Proses klasifikasi dengan seleksi fitur *Chi Square Statistic*

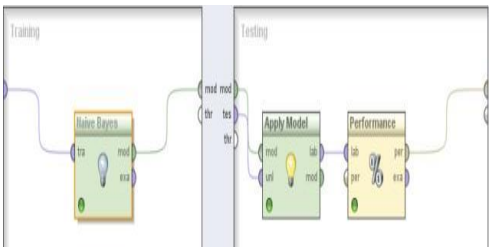
Read Excel digunakan untuk membaca *dataset* yang disimpan di dalam *file excel*. *Process Documents* merupakan tahap *Text Preprocessing* yaitu terdiri dari tahapan *transform case, tokenize*, dan *filter stopwords (dictionary). Weight by Chi* merupakan tahap penyeleksian fitur menggunakan *Chi Square Statistic.* Dan *Optimize Selection* dengan memilih parameter *forward selection. Forward selection* dimulai dengan tidak adanya fitur dan akan memulai menambahkan satu persatu fitur, sampai tidak ada lagi fitur yang mungkin dapat menurunkan *error* secara signifikan.

Cross-validation akan membagi data ke dalam dua *subset*, yang pertama untuk data *training* dan yang satunya lagi untuk data *testing*. Selanjutnya akan dilakukan pertukaran fungsi dari dua *subset* tersebut, sehingga *subset* yang sebelumnya sebagai data *training* akan menjadi data *testing* demikian sebaliknya. *X-validation* akan melakukan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian. Proses *X-validation* dilakukan untuk meningkatkan akurasi klasifikasi. Proses ini diletakan didalam proses *optimize selection*. Didalam proses *X-validation* terdapat proses klasifikasi yang dilakukan oleh algoritma *naïve bayes* seperti ditunjukan pada Gambar 5 *validation*.



Gambar 5 *Validation*

Proses klasifikasi disini adalah untuk menentukan sebuah kalimat terklasifikasi ke dalam kategori positif, negatif dan netral, berdasarkan nilai perhitungan probabilitas *naïve bayes* yang lebih besar. Misalnya jika hasil dari probabilitas kalimat untuk kategori negatif lebih besar daripada kategori positif dan netral maka kalimat tersebut termasuk ke dalam kategori negatif. Begitu juga sebaliknya dengan kategori positif dan netral.



Gambar 6 Klasifikasi *Naïve Bayes*

Hasil Pengujian *Naïve Bayes* dan *Chi Square Statistic*

Untuk mengetahui hasil pengujian *Naïve Bayes* dan *Chi Square Statistic* yaitu dengan menghitung tingkat akurasi pada tabel *Confussion Matrix*. Ditunjukan pada tabel 6 *Confussion Matrix Chi Square Statistic* dan *Naïve Bayes*.

Tabel 6 *Confussion Matrix Chi Square Statistic* dan *Naïve Bayes*

	***true* Netral**	***true* Negatif**	***true* Positif**
***pred.* Netral**	404	61	45
***pred.* Negatif**	6	71	1
***pred.* Positif**	4	3	45

Pada Tabel 6 diperoleh untuk jumlah *True* Netral (TNet) adalah 404, *True* Negatif (TN) adalah 71, *True* Positif (TP) adalah 45, *False* Netral (FNet) adalah 10, *False* Negatif (FN) adalah 64 dan *False* Positive (FP) adalah 46. Berdasarkan perhitungan tingkat akurasi dari data yang diperoleh pada tabel 6, menunjukan bahwa algoritma klasifikasi *Naïve Bayes Classifier* dan seleksi fitur *Chi Square Statistic* pada *tweets* bahasa Indonesia dan bahasa sunda mendapatkan akurasi sebesar 81.25%.

$$Accuracy = \frac{TP+TN+TNet}{TP+TN+FN+FP+FNet+TNet} \quad (10)$$

$$= \frac{45+71+404}{45+71+64+46+10+404}$$

$$= \frac{520}{640} \times 100\ \% = 81{,}25\ \%$$

KESIMPULAN

Algoritma klasifikasi *Naïve Bayes Classifier* dan metode seleksi fitur *Chi Square Statistic* dapat mengurangi fitur-fitur yang tidak relevan pada proses klasifikasi *Naïve Bayes Classifier* pada *tweets* bahasa Indonesia dan bahasa sunda.

Hasil pengujian terhadap penggabungan metode algoritma klasifikasi *Naïve Bayes* dan *Chi Square Statistic* pada

data uji mendapatkan akurasi sebesar 81.25%.

Pada penelitian berikutnya dapat dikembangkan dengan menggunakan metode klasifikasi lain seperti *K-Nearest Neighbor* (K-NN), *Neural Network*, *Support Vector Machine* dan lain-lain. Dan menggabungkan dengan metode pemilihan fitur yang lain, seperti *Genetic Algorithm*, *Term Frequency x Inverse Document Frequency*, *Gini Index* dan lain-lain.

DAFTAR PUSTAKA

[1] Routray, P., Swain, C. K. & Mishra, S.P., 2013. "A Survey on Sentiment Analysis. *International Journal of Computer Applications",* Agustus, 70(10), pp. 1-8

[2] Berry, M.W. & Kogan, J. 2010. "*Text Mining Aplication and theory"*. WILEY : United Kingdom.

[3] Yang, Y., & Pedersen, J. O. 1997. "A comparative study on feature selection in text categorization". ICML, (hal. 412--420).

[4] Gorunescu, F. 2011. "*Data Mining Concepts, Model and Techniques"*. Berlin: Springer.

[5] Feldman, R & Sanger, J. 2007. "The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data". Cambridge University Press : New York.